DXE_EMTR 2021
Second assignment (20% of total grade)

Please submit the assignment by 9 Dec in the IS MUNI system. You are allowed and encouraged to work in groups of maximum size 3. Please don't forget to submit your R-code too.

1 Causal graphical models (5%)

We are interested in quantifying the effect of discrimination on wages. Assume that gender influences the discrimination but has no direct impact on both occupation and wages. Furthermore, we assume that discrimination influences both occupation and wages. Occupation also have a direct effect on wages. Also, unobserved ability influences both occupation and wages.

- Plot the causal graph (denote it as G_1).
- List all the causal and non-causal paths from discrimination to wages.
- Determine if it is possible to identify the causal effect of discrimination on wages based on observed probability distribution of gender, discrimination, occupation and wages.
- Discuss the pros and cons of adjusting for the occupation.

Now assume that some occupations are preferred by different genders and that there is a direct link from gender to occupation.

- Plot the causal graph (denote it as G_2).
- List all the causal and non-causal paths from discrimination to wages.
- Determine if it is possible to identify the causal effect of discrimination on wages based on observed probability distribution of gender, discrimination, occupation and wages.
- Discuss the pros and cons of adjusting for the occupation and/or gender.

Now furthermore assume that we have a good proxy variable for ability (e.g. composite IQ and EQ score), so that ability is now observed.

- Plot the causal graph (denote it as G_3).
- List all the causal and non-causal paths from discrimination to wages.
- Determine if it is possible to identify the causal effect of discrimination on wages based on observed probability distribution of gender, discrimination, occupation and wages.
- Discuss the pros and cons of adjusting for the occupation and/or gender and/or ability.
- Assume that discrimination variable is binary and propose an estimator for (total) average treatment effect of discrimination on wages.
- Complement your analysis of G_3 with a small simulation study. Convince the reader that the estimator that you proposed is able to recover the true (total) average treatment effect of discrimination on wages. Make any reasonable simplifications in order to illustrate your point.

You may find `dagitty` helpful for these tasks (either in `R` or at `www.dagitty.net`).

2 Selection on Observables (7%)

Consider the following hypothetical situation. We have a population of 10 individuals and we happen to know both their contrafactual earnings: if they go to job training programme ($D = 1$) they will receive $Y(1)$. Or if they do not participate in the job training programme ($D = 0$) they will get $Y(0)$. We also have information on the different background they have: some individuals come from cities ($X = 1$) while others from rural areas ($X = 0$).

Unit	$Y(1)$	$Y(0)$	D	X
1	20	19	1	1
2	23	21	1	1
3	36	20	0	1
4	43	19	0	1
5	19	19	0	1
6	23	55	0	0
7	26	41	1	0
8	21	21	1	0
9	33	37	0	0
10	16	17	1	0

(1a) What is the true average treatment effect (ATE) of job training programme on earnings?

(1b) What is the true average treatment effect on the treated (ATT) of job training programme on earnings?

(1b) What is the true average treatment effect on the untreated (ATU) of job training programme on earnings?

(1d) What is the true average treatment effect on the treated of job training programme on earnings for those from cities?

(1e) What is the true average treatment effect on the treated of job training programme on earnings for those from rural areas?

(1f) Suppose that you can persuade every person to either attend the training or not based on what is more beneficial for her/him in terms of earnings. What would be the average effect of such ideal intervention?

This information is, of course, not available to the outside analyst, who only observes (Y, D, X), where $Y = Y(1) \cdot D + (1 - D) \cdot Y(0)$ is the observed earnings.

(2a) Calculate unadjusted differences in mean observed outcomes job training participants $(D = 1)$ and non-participants $(D = 0)$.

(2b) During the lecture, we have seen the following decomposition:

$$\begin{aligned} E[Y|D=1] - E[Y|D=0] &= \overbrace{\underbrace{E[Y(1)|D=1]}_{E[Y|D=1]} - \underbrace{E[Y(0)|D=1]}_{\text{unobserved}}}^{ATT=E[Y(1)-Y(0)|D=1]} \\ &+ \underbrace{\underbrace{E[Y(0)|D=1]}_{\text{unobserved}} - \underbrace{E[Y(0)|D=0]}_{E[Y|D=0]}}_{\text{Selection bias}} \end{aligned}$$

Calculate these quantities and confirm that this equation holds for our 10 individuals.

(2c) We have also seen this way of decomposing the observe differences $E[Y|D = 1] - E[Y|D = 0]$:[1]

$$\begin{aligned} E[Y|D=1] - E[Y|D=0] &= \overbrace{\underbrace{E[Y(1)]}_{\text{unobserved}} - \underbrace{E[Y(0)]}_{\text{unobserved}}}^{ATE=E[Y(1)-Y(0)]} \\ &+ \underbrace{\underbrace{E[Y(0)|D=1]}_{\text{unobserved}} - \underbrace{E[Y(0)|D=0]}_{E[Y|D=0]}}_{\text{Selection bias}} \\ &+ \underbrace{Pr(D=0)\left(\overbrace{E[Y(1)-Y(0)|D=1]}^{ATT} - \overbrace{E[Y(1)-Y(0)|D=0]}^{ATU}\right)}_{\text{Heterogenous treatment effects bias}} \end{aligned}$$

[1] This derivation can be found on page 132 of Cunningham, Scott. Causal Inference. Yale University Press, 2021. The author made the book available for free here: `https://mixtape.scunning.com/potential-outcomes.html`. Reading the whole chapter could be helpful for this exercise.

Calculate these quantities and confirm that this equation holds for our 10 individuals.

(2d) Based on all these results, discuss if selection-on-observables was a valid assumption to make.

```
dat <- matrix(c(20,19,1,1,
                23,21,1,1,
                36,20,0,1,
                43,19,0,1,
                19,19,0,1,
                23,55,0,0,
                26,41,1,0,
                21,21,1,0,
                33,37,0,0,
                16,17,1,0),
                byrow=T, ncol=4)
colnames(dat) <- c("Y1","Y0","D","X")
rownames(dat) <- c(1,2,3,4,5,6,7,8,9,10)
dat <- as.data.frame(dat)
```

3 Replication (8%)

Here are data archives of Joshua Angrist `https://economics.mit.edu/faculty/angrist/data1/data` and of Daron Acemoglu `https://economics.mit.edu/faculty/acemoglu/data`. There are datasets and replication files in STATA (another program that many economists use for conducting statistical analyses).[2]

- Choose one paper and read it. Preferably choose a paper that is close to your research agenda, if this is possible.
- Replicate the main results in `R`. You don't have to replicate all the results (e.g. those in appendices etc) in the paper, only the most interesting set of main results. R-package `haven` may be useful for you for loading the datasets into `R`.
- Explore if/how these results are sensitive to the functional form specification of the regressions or other model choices. These are some examples you may/may not consider:
  - check whether adding a quadratic term of certain covariate into a regression changes the results substantially,
  - check whether adding a relevant interaction term changes the results substantially,
  - whether having the outcome variable in logarithm (or some other transformation) leads to very different conclusions,
  - look if the results hold if you only look at a particular subsample,
  - be curious and critical.

  There are many modifications you may consider. The changes that you suggest should be motivated by some economic reasoning, they should not be completely artificial.
- Write down what you found: explain what did you try and motivate why. Discuss any potentially interesting findings.

Make sure to comment your code and make your best effort to adhere to some reasonable coding standards. Your code must be easy to read and it should take minimal effort to reproduce your results. Present your results in a coherent way and whenever possible make use of visualization.

[2] Please do not choose: Acemoglu, Daron, Simon Johnson, and James A. Robinson. 'The colonial origins of comparative development: An empirical investigation.' American economic review 91.5 (2001): 1369-1401. which is replicated in the `lecture4.R` in the IS MUNI system. We will finish the Instrumental variables topic at the beginning of the next session.